Continuous Iterative Guided Spectral Class Rejection Classification Algorithm: Part 1

Rhonda D. Phillips *Student Member, IEEE*, Layne T. Watson *Fellow, IEEE*,
Randolph H. Wynne *Member, IEEE*, and Naren Ramakrishnan *Member, IEEE*

Abstract—This paper outlines the changes necessary to convert the iterative guided spectral class rejection (IGSCR) classification algorithm to a soft classification algorithm. IGSCR uses a hypothesis test to select clusters to use in classification and iteratively refines clusters not yet selected for classification. Both steps assume that cluster and class memberships are crisp (either zero or one). In order to make soft cluster and class assignments (between zero and one), a new hypothesis test and iterative refinement technique are introduced that are suitable for soft clusters. The new hypothesis test, called the (class) association significance test, is based on the normal distribution, and a proof is supplied to show that the assumption of normality is reasonable. Soft clusters are iteratively refined by creating new clusters using information contained in a targeted soft cluster. Soft cluster evaluation and refinement can then be combined to form a soft classification algorithm, continuous iterative guided spectral class rejection (CIGSCR).

I. Introduction

The classification of remotely sensed imagery is essential for many remote sensing applications such as natural resource management, change detection, species identification, etc. Crisp classifications assign each pixel or sample to one class in the particular classification scheme, which can be interpreted as picking the class that has the highest probability of containing the sample (when probability models are used for classification). Alternatively, soft classifications contain information on possible memberships in multiple classes, not just the most likely class. Soft or subpixel classifications are of considerable interest in the remote sensing community as this type of classification can effectively model geographic data whose natural boundaries rarely coincide with pixel boundaries. Furthermore, pixels can also contain multiple species that are commingled, leading to classification difficulty. Individual classes within the classification scheme can have overlapping electromagnetic reflectance spectra, making it difficult to discriminate between these classes. Scientists have successfully used soft classification for applications such as land cover mapping [1], vegetation mapping [2], and the classification of snow [3], to name a few. Popular methods for obtaining soft classifications of remotely sensed images include fuzzy c-means [4] and spectral unmixing [5].

Semisupervised classification has received a good deal of attention in the remote sensing community as remote sensing datasets are characterized by a large number of dimensions (hyperspectral imagery) and limited training data. While training data is expensive to obtain in any discipline, it is especially so in remote sensing as the labeling of image data typically requires extensive knowledge of the study area, multiple data sources, and/or physically visiting the study area to identify classes. Semisupervised learning can be used to supplement a labeled training set with unlabeled data to mitigate the Hughes phenomenon (overfitting of a classification when the training data is insufficient for the number of dimensions present in the dataset to be classified) [6][7].

Semisupervised classification algorithms such as the iterative guided spectral class rejection (IGSCR) algorithm ([8],[9],[10]) have the additional benefit of providing a high level of automation compared to strictly supervised classification algorithms. In remote sensing, informational class categories that make up a classification scheme are defined prior to classification and are identified by humans, whereas spectral classes or clusters have mathematical properties (such as mathematically homogeneous spectral waveforms) and are more difficult for humans to identify. For example, suppose a forest/nonforest classification is desired, and forest and nonforest are the informational class categories. Each informational class is composed of multiple spectral classes that can be used in supervised classification, and the individual spectral classes may not be spectrally similar to each other despite all being part of one informational class. Consider the wide range of tree species that could potentially make up a forest informational class in a particular image. An unsupervised technique such as clustering can identify individual classes that

R.D. Phillips, L.T. Watson, and N. Ramakrishnan are with the Departments of Computer Science and Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

R.H. Wynne is with the Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

are mathematically homogeneous, and has the additional property of guaranteeing that all types of land cover present in a dataset are represented in the spectral classes (clusters). Both tasks are nontrivial for humans to perform when identifying spectral classes for supervised classification. Therefore semisupervised classification algorithms that involve clustering can automatically identify and label spectral classes, providing significant automation over supervised or unsupervised classification alone.

The purpose of this work is to develop a semisupervised soft clustering framework, analogous to the framework in IGSCR, that is capable of producing soft classifications of remotely sensed images. This framework will potentially affect semisupervised classification algorithms that have labeled data and involve clustering. Soft clustering retains all information regarding the proximity of data points to clusters, and will therefore directly produce a soft classification and will potentially provide better training spectral classes for a supervised decision rule. The major challenges to converting the discrete IGSCR to a fully continuous algorithm producing soft classification are in converting the underlying inherently discrete models and algorithms to suitable continuous models and algorithms while preserving the automated spectral class identification properties of IGSCR. More specifically, a hypothesis test that is fundamental to IGSCR is based on the discrete binomial probability distribution. A hypothesis test based on a new continuous probability distribution is necessary in continuous IGSCR (CIGSCR). IGSCR uses an iterative cluster refinement framework that breaks down under soft clustering, and therefore a new iterative cluster refinement method is developed for CIGSCR. Furthermore, soft clustering allows for the magnification of distances using radial functions that changes soft clusters but would have no effect on hard clusters.

The remainder of the paper is organized as follows. Section II reviews related literature on semisupervised learning and semisupervised clustering. Section III describes IGSCR in detail, and Section IV describes CIGSCR. Section V introduces the association significance test, a hypothesis test based on a new distribution that will be suitable for evaluating the class associations to soft clusters. Section VI discusses changes necessary for the iterative refinement of soft clusters. Section VII concludes the paper.

II. Background

Semisupervised learning occurs when unlabeled data are used in addition to labeled data to produce a classification [11]. Semisupervised learning can be more accurate than supervised learning (for a given set of labeled data) if knowledge of the underlying data distribution $p(x)$ (gained through the unlabeled data) contributes to knowledge of the conditional distribution $p(c|x)$, where c is the class label for data point x [11]. When assumptions about $p(c|x)$ are incorrect, using information about $p(x)$ can actually degrade classification accuracy [11].

An assumption commonly used in semisupervised learning is that if two particular points in a dense region are "close," their corresponding class labels should also be "close." In the context of clustering, this indicates that two points contained in the same cluster are likely to be in the same class, which is known as the "cluster assumption" [12]. Semisupervised learning methods that invoke the cluster assumption include the method proposed in [13]. Unfortunately, this assumption is sometimes not true as clusters are not necessarily composed of one class. Several clustering methods have been suggested that seek to form clusters based on both traditional clustering criteria and secondary criteria that could include a correlation between clusters and classes. Clustering methods that use additional information to influence clusters are known as *semisupervised clustering* (distinct from semisupervised learning) methods.

One method that seeks to influence the formation of clusters is clustering with constraints. In these methods, constraints are provided in the form of must-link constraints where two samples should appear in the same cluster and cannot-link constraints where two samples should not appear in the same cluster. These constraints are used with a traditional clustering method such as k-means, and the constraints can be strictly enforced algorithmically [14] or by using a modified objective function [15]. When using an objective function, there is no guarantee that all constraints will be satisfied. Basu et al. [16] suggested a method by which constraints that are informative can be selected and used in clustering, and Bilenko et al. [17] used constraints to learn a distance metric that would provide a good clustering. Halkidi et al. [18] use constraints to measure the quality of a clustering and tune Euclidean distance weight parameters to find the "best" clustering. Bouchachia and Pedryz [19] introduced a soft semisupervised clustering method with an objective function that accounts for prior information in the form of class labels. Having class labels can be viewed as a special case of having constraints as must-link and cannot-link constraints can be generated from the labeled data. Other methods that use additional information to form clusters include information bottleneck ([20], [21], [22]) and discriminative clustering [23]. These algorithms form a clustering objective function that measures distortion of the auxiliary data due to clustering.

Semisupervised learning has been used in the remote sensing community for some time to supplement limited training samples in the classification of remotely sensed images. The application of semisupervised learning to correct classification overfitting was studied in [7]. Jeon and Landgrebe used semisupervised techniques (including clustering) to perform classifications on entire images when only one class is of interest and labeled [24]. Multiple semisupervised methods based on support vector

machines (SVM) have been developed for the classification of hyperspectral imagery ([25], [26]), and Gòmez-Chova et al. used clustering and SVMs to form a semisupervised classification method [27]. IGSCR also utilizes clustering in a semisupervised framework to classify remotely sensed images ([8], [9], [10]). Due to its high accuracy and automation, IGSCR is a frequently used hybrid classification method in the remote sensing community ([28], [29], [30], [31]).

III. IGSCR

IGSCR is a classification method that uses clustering to generate a classification model $p(c_i|x)$ where x is a multivariate sample to be classified and c_i, $i = 1, \ldots, C$, is the ith class where there are C classes in the classification scheme. IGSCR uses clustering to estimate $p(k_j|x)$ in the expression

$$p(c_i|x) = \sum_{j=1}^{K} p(c_i, k_j|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x), \tag{1}$$

where k_j, $j = 1, \ldots, K$, is the jth cluster out of K total clusters. IGSCR also uses the clusters to train a decision rule using Bayes' theorem [32]

$$p(k_j|x) = \frac{p(x|k_j)p(k_j)}{\sum_{i=1}^{K} p(x|k_i)p(k_i)}. \tag{2}$$

The prior probabilities of the clusters $p(k_j)$ are assumed to be equal.

Clustering is performed using a discrete clustering method such as k-means that minimizes the objective function

$$J(\rho) = \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}\rho_{ij} \tag{3}$$

subject to

$$\sum_{j=1}^{K} w_{ij} = 1$$

where $w_{ij} \in \{0, 1\}$ is the value in the ith row and jth column of the partition matrix $W \in \Re^{n\times K}$, $U^{(j)} \in \Re^B$ is the prototype for the jth cluster k_j, $x^{(i)} \in \Re^B$ is the ith data point, and $\rho_{ij} = ||x^{(i)} - U^{(j)}||_2^2$. The clusters $k_1, \ldots, k_K$ form a partition of $\{x^{(i)}\}_{i=1}^n$. The algorithm for k-means requires K initial cluster prototypes and iteratively assigns each sample to the closest cluster using

$$w_{ij} = \begin{cases} 1, & \text{if } j = \underset{1\le j\le K}{\operatorname{argmin}}\, \rho_{ij}, \\ 0, & \text{otherwise,} \end{cases}$$

followed by the cluster prototype (mean) recalculation

$$U^{(j)} = \sum_{i=1}^{n}(w_{ij}x^{(i)}) \Big/ \sum_{i=1}^{n} w_{ij}$$

once W has been calculated [33]. This process, guaranteed to terminate in a finite number of iterations, continues until no further improvement is possible, terminating at a local minimum point of (3).

IGSCR uses labeled data in a semisupervised clustering framework to locate clusters that map to classes in a given classification scheme. IGSCR requires a labeled set of training data comprised of individual samples within the image to be classified and corresponding class labels. Rather than using the labeled data to train a decision rule directly, the entire image is clustered, thereby capturing the inherent structure of all the data and not just the labeled samples. The clusters represent spectral classes, and in remote sensing, each spectral class ideally maps to exactly one class in the final classification scheme. Once clusters are generated, each cluster must be mapped to one class or rejected as impure. While theoretically each cluster should contain samples belonging to only one informational class, in practice clusters (spectral classes) that contain predominantly samples of one class can contain a few samples from other classes because of inherent errors. However, if a cluster contains too many samples from different classes, the cluster itself is considered confused and should not be labeled with one class. Impure clusters are rejected and can be further refined in the iterative part of the algorithm.

The test for cluster purity is performed using the labeled training set. IGSCR produces a hard classification and uses a discrete clustering method where each sample is assigned to exactly one cluster. Let $V_{c,j}$ be the binomial random variable denoting the number of labeled samples assigned to the jth cluster that are labeled with a particular cth class. Let p be the user-supplied cluster homogeneity threshold ($p = .9$ would indicate a cluster is 90% pure with respect to the majority class), and let α be the user-supplied acceptable one-sided Type-I error for a statistical hypothesis test. Then if c is the majority class represented in the jth cluster, the jth cluster is rejected if $P(Z < \hat{z}) < 1 - \alpha$ where Z is a standard normal random variable, m is the number of labeled samples in the jth cluster, and

$$\hat{z} = \frac{v_{c,j} - mp}{\sqrt{mp(1-p)}}. \tag{4}$$

(Typically a continuity correction of 0.5 is added in the numerator of (4).)

If a cluster is rejected, the samples making up that cluster can be reclustered in subsequent iterations. All samples belonging to pure clusters are removed from the image being clustered, resulting in only samples belonging to impure clusters being reclustered. Once more clusters are generated, those clusters are evaluated for purity, removed from the image, and clustering is performed again until termination criteria are met. All samples can belong to pure clusters, leaving no remaining samples to be clustered, no pure clusters could be found in the previous iteration, meaning that the clustering would continue to be performed on the same data, resulting in the same impure clusters (assuming deterministic cluster seeding), or a set number of iterations can be reached, resulting in termination of the iteration. Note that deterministic seeding ensures that the iteration will terminate, even without specifying a maximum number of iterations.

Once the iterative clustering is complete, one or more classifications is performed. The first classification is called the iterative stacked (IS) classification because it is the result of combining or "stacking" all cluster assignments over all iterations (each sample will be assigned to at most one accepted cluster). Assume that all samples not assigned to an accepted cluster are combined to form one cluster k_{K+1}, and the class assignment for that cluster is "unclassified" or c_{C+1}. Then the IS assignment for a pixel using (1) is

$$\text{IS}(x) = \underset{1 \le i \le C+1}{\text{argmax}}\, p(c_i|x) = \underset{1 \le i \le C+1}{\text{argmax}} \sum_{j=1}^{K+1} p(c_i|k_j, x)p(k_j|x),$$

where

$$p(c_i|k_j, x) = \begin{cases} 1, & \text{if } k_j \text{ is labeled } c_i, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$p(k_j|x) = \begin{cases} 1, & \text{if } x \in k_j, \\ 0, & \text{otherwise,} \end{cases}$$

since cluster assignments are discrete.

The second possible classification, the decision rule (DR) classification, uses the pure clusters to form a decision rule. Recall in (2) that

$$p(k_j|x) = \frac{p(x|k_j)}{\sum_{i=1}^{K} p(x|k_i)}$$

when all the $p(k_j)$ are equal. Traditionally, the maximum likelihood decision rule, assuming a multivariate normal distribution

$$p(x|k_j) = 2\pi^{-B/2}|\Sigma_j|^{-1/2} e^{-\frac{1}{2}(x-U^{(j)})^T \Sigma_j^{-1}(x-U^{(j)})},$$

is used where Σ_j is the covariance matrix of the jth cluster [34]. Since IGSCR produces hard classifications, the full probability need not be calculated as determining only the cluster associated with the maximum probability is necessary. The DR classification function is

$$\text{DR}(x) = \underset{1 \le i \le C}{\text{argmax}}\, p(c_i|x) = \underset{1 \le i \le C}{\text{argmax}} \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x), \tag{5}$$

where

$$p(k_j|x) = \begin{cases} 1, & \text{if } j = \underset{1 \le j \le K}{\text{argmax}}\big(-\ln|\Sigma_j| - (x - U^{(j)})^T \Sigma_j^{-1}(x - U^{(j)})\big), \\ 0, & \text{otherwise.} \end{cases}$$

A final classification, the iterative stacked plus (IS+) classification, combines the DR and IS classifications. If a sample is labeled as unclassified in the IS classification, the DR class value is used for the IS+ classification, otherwise the IS class value is used for that particular sample. The IS+ classification function is

$$\text{IS+}(x) = \begin{cases} \text{IS}(x), & \text{if } x \notin k_{K+1}, \\ \text{DR}(x), & \text{otherwise.} \end{cases}$$

IV. CIGSCR

Continuous IGSCR (CIGSCR) uses a similar semisupervised clustering framework to the one established in IGSCR to produce a soft or probabilistic classification instead of a hard classification, and uses continuous algorithms and models instead of discrete algorithms and models. Recall in (1) that $p(c_i|k_j, x)$ and $p(k_j|x)$ are either 0 or 1 (discrete) in practice in IGSCR. $p(c_i|k_j, x)$ is necessarily discrete because while several clusters can comprise one class, only one class (theoretically) can label the members of a particular cluster, but there are no similar restrictions on $p(k_j|x)$. In fact, the clustering algorithm and the maximum likelihood decision rule indicate positive probabilities that a sample is associated with each cluster, but IGSCR makes an assignment only to the cluster with the highest probability.

Consider a soft clustering algorithm that minimizes the objective function [35]

$$\begin{aligned} J(\rho) = \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}^{p}\rho_{ij} \quad & \text{subject to} \\ \sum_{j=1}^{K} w_{ij} = 1 \text{ for each } i & \end{aligned} \tag{6}$$

where $w_{ij} \in (0, 1)$ is the value in the ith row and jth column of the weight matrix $W \in \Re^{n\times K}$ (analogous to the partition matrix W in (3)), $U^{(j)} \in \Re^{B}$ is the jth cluster prototype, $p > 1$, and $\rho_{ij} = \rho(x^{(i)}, U^{(j)}) = ||x^{(i)} - U^{(j)}||_2^2$ is the Euclidean distance squared. The algorithm that minimizes this objective function is similar to that of k-means in that it first calculates

$$w_{ij} = \frac{(1/\rho_{ij})^{1/(p-1)}}{\sum_{k=1}^{K}(1/\rho_{ik})^{1/(p-1)}}$$

for all i and j followed by calculating updated cluster prototypes

$$U^{(j)} = \sum_{i=1}^{n} w_{ij}^{p} x^{(i)} \Big/ \sum_{i=1}^{n} w_{ij}^{p}.$$

This iteration (recalculation of the weights followed by recalculation of cluster prototypes, following by recalculation of the weights, etc.) is guaranteed to converge (with these definitions of ρ_{ij}, $U^{(j)}$, and w_{ij}) for $p > 1$ [36].

With a continuous alternative to the discrete hypothesis test and a continuous alternative to the IGSCR iterative cluster refinement that follows in Sections 5 and 6, the classification function for IS classification is

$$\text{IS}(x) = p(c_i|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x), \tag{7}$$

where $p(k_j|x)$ is estimated using w_{ij} and $p(c_i|k_j, x)$ does not change from IGSCR. The classification function for the DR classification is

$$\begin{aligned} \text{DR}(x) = p(c_i|x) &= \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x) \\ &= \frac{\sum_{j=1}^{K} p(c_i|k_j, x)\left[\dfrac{2e^{-\frac{1}{2}(x-U^{(j)})^T\Sigma_j^{-1}(x-U^{(j)})}}{\pi^{B/2}|\Sigma_j|^{1/2}}\right]}{\sum_{l=1}^{K}\left[\dfrac{2e^{-\frac{1}{2}(x-U^{(l)})^T\Sigma_l^{-1}(x-U^{(l)})}}{\pi^{B/2}|\Sigma_l|^{1/2}}\right]}. \end{aligned} \tag{8}$$

An analog for the IS+ classification is unnecessary in CIGSCR as all samples will be part of pure clusters and will be classified.

V. Association Significance Test

A key component in the IGSCR semisupervised clustering framework is the homogeneity test used to determine if a cluster contains a statistically significant proportion of one class. This test provides a basis for rejecting a cluster for further refinement, the second phase of the semisupervised clustering.

A cluster might be composed of more than one class because the cluster itself is in fact composed of more than one cluster. A cluster might also contain more than one class because the initial clusters were determined in such a way as to prevent a cluster from moving toward a particular class. It would be useful to determine which clusters are not spectrally pure (contain more than one class with high probability) so that the cluster can be further refined, and if no refinement is possible (any number of iteration ending criteria are met), the cluster should not be used in the classification model. Statistical hypothesis tests provide a mechanism for determining class purity once an appropriate statistical model is selected for the data.

In hard IGSCR with hard clustering, the notion of a pure cluster is clear. Each sample will belong to one and only one cluster. A cluster can be 100% homogeneous when all labeled samples contained within that cluster belong to only one class. Although this is possible, it is unlikely that one cluster contains only one class because of inherent error in the labeling process and because two different informational class categories can contain spectrally similar samples. Once a homogeneity level is determined, a rigorous hypothesis test can be applied to select clusters that contain a certain percentage of one class, with that percentage unlikely to be observed in a particular cluster randomly.

Using soft clusters introduces complications to assessing and determining cluster purity. The first question might be whether a soft cluster can be spectrally pure, because being soft might indicate that clusters are naturally comprised of multiple classes. However, just as the goal in IGSCR is to determine clusters that are representative of just one predominant class, that goal holds in CIGSCR with soft clusters. Soft clusters are composed of different portions of each sample or pixel within an image, meaning that each sample has a positive probability of being in different individual classes or clusters. When samples labeled with different classes have a positive probability of belonging to the same cluster, that does not indicate that the cluster really contains two different classes, but rather perhaps that while the pixels have strong associations with different classes, there is also a positive (although possibly small) probability that each pixel actually belongs to or partially belongs to the majority class within the cluster. Both cases (the cluster is confused or the cluster is not confused but the pixels labeled with different classes still have small associations with the same class) are possible in soft clustering. The appropriate test for soft clusters is not which pixels "belong" to a particular cluster (they all "belong" to some degree), rather how strongly pixels from different classes belong to a particular cluster. If pixels from only one class have strong associations with a cluster when compared to pixels labeled with other classes, then the cluster should be labeled with that most strongly associated class. In this manner, each pixel/sample is associated by varying degrees with multiple spectrally pure clusters that are mapped to individual classes, ultimately producing a soft classification output when each sample is then mapped to different individual classes with varying probabilities.

A. Distribution

Developing a hypothesis test to assess purity of clusters requires a random variable and knowledge of the distribution of that random variable. In IGSCR, a cluster can be considered pure and labeled with a class if the number of labeled samples belonging to the class is high compared to the number of labeled samples not belonging to the class. The random variable of interest, $V_{c,j} = \sum_{i \in I_j} V_{ic}$, is the count of the number of labeled samples belonging to the cth class for a particular jth cluster where i is the pixel index, I_j is the index set of labeled pixels in the jth cluster, and V_{ic} is the Bernoulli random variable corresponding to the ith pixel being associated with the cth class. A hypothesis test can be developed using the binomial distribution, or the less computationally intensive normal distribution, which approximates the binomial distribution well when the number of labeled samples is large.

In CIGSCR, the random variable and distribution are more complicated as there are class memberships (either 0 or 1) and cluster memberships (between 0 and 1). Building a test on only the class memberships is not useful as each labeled sample will have some positive probability of belonging to a particular cluster, making the results of the test the same for each cluster unless memberships are also considered. In this case, the association of a sample to a particular class (the majority class, for example) is still a Bernoulli trial. Each pixel also has a weight vector, $w_{i\cdot}$, indicating the probability of membership to each cluster. The random variable of interest is the sum of the memberships for the cth class and weights to the jth cluster,

$$Y_{c,j} = V_{1c}W_{1j} + V_{2c}W_{2j} + \cdots + V_{nc}W_{nj},$$

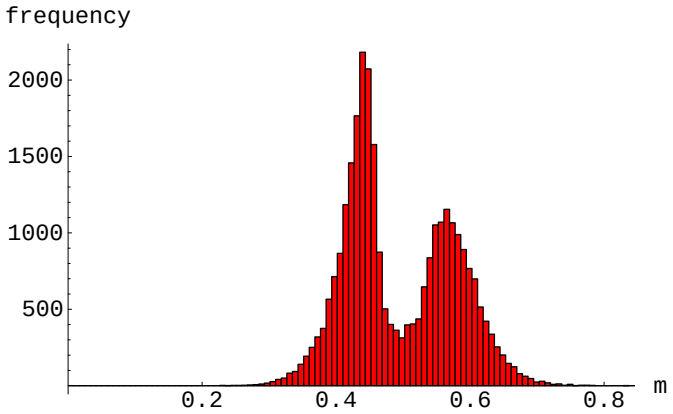

Fig. 1. Histogram of cluster weights in one cluster, K=2

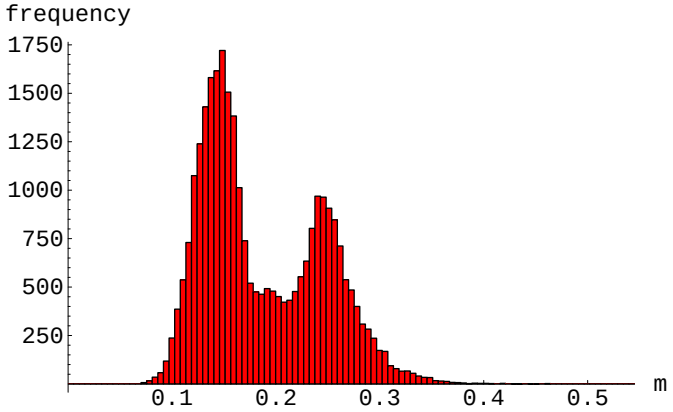

Fig. 2. Histogram of cluster weights in one cluster, K=5

where n is the total number of labeled samples. The labels of the classified pixels are independent of cluster assignment, making an assumption that V_{ic} and W_{ij} are independent reasonable. Furthermore, the training samples are labeled prior to clustering, making the random variable of interest

$$Y_{c,j}|(V_{1c}, V_{2c}, \ldots, V_{nc}) = \sum_{i=1}^{n} W_{ij}\delta_{\phi(i),c},$$

where $\phi(i)$ is the label of the ith pixel, and

$$\delta_{\phi(i),c} = \begin{cases} 0 \text{ if } \phi(i) \neq c, \\ 1 \text{ if } \phi(i) = c, \end{cases}$$

is the Kronecker delta. The probability density function (pdf) of $Y_{c,j}|(V_{ic}, i = 1, \ldots, n) = \sum_{i=1}^{n} W_{ij}\delta_{\phi(i),c}$ is the pdf of a sum of individual cluster weights.

Figs. 1 and 2 contain experimental frequency histograms of weights w_{ij} for two clusters ($K = 2$) of a satellite image. The distribution of the cluster weights appears to be multimodal, which is consistent with the data having multiple inherent classes, indicating that W_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, K$ would not be identically distributed. A closed form distribution is not readily available for W_{ij}, but a closed form distribution, or at least a reasonable approximate closed form distribution, for $W_{+j} = \sum_{i=1}^{n} W_{ij}$ exists.

B. Normal Approximation to $Y_{c,j}$

Suppose an image x contains n pixels $x^{(i)} \in \Re^B$, $i = 1, \ldots, n$. For K fixed cluster centers $U^{(k)} \in \Re^B$, $k = 1, \ldots, K$, the

assigned weight of the ith pixel to the jth cluster is

$$w_{ij} = \frac{1/||x^{(i)} - U^{(j)}||_2^2}{1 \Big/ \sum_{k=1}^{K} ||x^{(i)} - U^{(k)}||_2^2},$$

which is the inverse of the distance squared over the sum of the inverse squared distances. (Such inverse distance weights are widely used, e.g., by Shepard's algorithm for sparse data interpolation.) Note this is the specific case in the soft clustering algorithm described above when $p = 2$. In this case where a remotely sensed image is to be clustered, it is reasonable to assume that $x^{(i)}$, $i = 1, \ldots, n$ are generated from a finite number of multivariate normal distributions. The act of clustering assumes that the data are generated from a finite number of distributions, and remotely sensed earth data are assumed to be generated from normal distributions. The following proof demonstrates that under these assumptions (pixels are generated from a finite number of normal distributions), the Lindeberg condition is satisfied and therefore the central limit theorem applies to the sum of a sequence of cluster weight random variables $\sum_{i=1}^{n} W_{ij}$. Let $q = \psi(i)$ denote the distribution from which $X^{(i)}$ was sampled.

Theorem: Let $X^{(i)}$, $i = 1, 2, \ldots$, be B-dimensional random vectors having one of Q distinct multivariate normal distributions. For $i = 1, 2, \ldots$ and $j = 1, \ldots, K$ define the random variables

$$W_{ij} = W_j(X^{(i)}) = \frac{1/||X^{(i)} - U^{(j)}||_2^2}{\sum_{k=1}^{K} 1/||X^{(i)} - U^{(k)}||_2^2},$$

where K is the number of clusters and $U^{(k)} \in \Re^B$ is the kth cluster center (and is considered fixed for weight calculation). Then for any $j = 1, \ldots, K$,

$$P\left\{ \frac{1}{B_{nj}} \sum_{i=1}^{n} (W_{ij} - a_{ij}) < x \right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$$

as $n \to \infty$, where $a_{ij} = \mathrm{E}[W_{ij}]$, $b_{ij}^2 = \mathrm{Var}[W_{ij}]$, and $B_{nj}^2 = \sum_{i=1}^{n} b_{ij}^2$.

Proof. W_{ij} is a bounded ($0 \le W_{ij} \le 1$) measurable function of a normal random variable, and is therefore a random variable with finite mean and variance. Fix j for the remainder of the proof, and let $q = \psi(i)$ denote which of the Q distributions $X^{(i)}$ is from. In order to prove

$$P\left\{ \frac{1}{B_{nj}} \sum_{i=1}^{n} (W_{ij} - a_{ij}) < x \right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz,$$

it is sufficient to verify the Lindeberg condition [32]:

$$\lim_{n\to\infty} \frac{1}{B_{nj}^2} \sum_{i=1}^{n} \int_{|x-a_{ij}|>\tau B_{nj}} (x - a_{ij})^2 dF_{\psi(i),j}(x) = 0,$$

for any constant $\tau > 0$ where $F_{\psi(i),j}(x)$ is the cumulative distribution function for W_{ij}.

For each q, $1 \le q \le Q$, define $I_q = \psi^{-1}(q) = \{i \mid \psi(i) = q, 1 \le i \le n\}$, $n_q = |I_q|$, and for $i \in I_q$ let $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ and $\mathrm{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$. Now considering only the independent and identically distributed random variables W_{ij}, $i \in I_q$, the Lindeberg condition holds:

$$\lim_{n_q\to\infty} \frac{1}{n_q \beta_{qj}^2} \sum_{i\in I_q} \int_{|x-\alpha_{qj}|>\tau\sqrt{n_q}\beta_{qj}} (x - \alpha_{qj})^2 dF_{qj}(x)$$

$$= \lim_{n_q\to\infty} \frac{1}{\beta_{qj}^2} \int_{|x-\alpha_{qj}|>\tau\sqrt{n_q}\beta_{qj}} (x - \alpha_{qj})^2 dF_{qj}(x) = 0.$$

Since β_{qj} is positive and finite, and the integral is finite, the limit of the integral is zero as $\sqrt{n_q}\beta_{qj} \to \infty$.

$W_{ij}, i = 1, 2, \ldots,$ are random variables from Q iid distributions, $F_{qj}, q = 1, \ldots, Q$, where the mean of the qth distribution is α_{qj}, the variance is β_{qj}^2, and the number of random variables from that distribution is n_q, where $\sum_{q=1}^{Q} n_q = n$. As $n \to \infty$ there

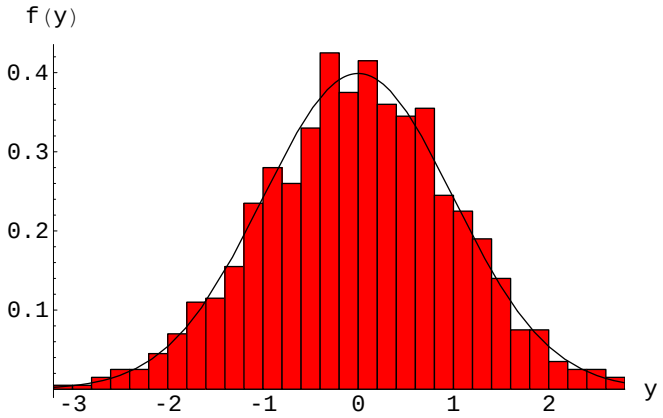

Fig. 3. Pdf of Y (with sample mean subtracted and divided by the standard deviation) compared to a standard normal distribution.

is at least one q for which $n_q \to \infty$. For this sequence of independent random variables from Q distributions, the Lindeberg condition is

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{B_{nj}^2} \sum_{i=1}^{n} \int_{|x-a_{ij}|>\tau B_{nj}} (x-a_{ij})^2 dF_{\psi(i),j}(x) \\
&= \lim_{n\to\infty} \frac{1}{\sum_{k=1}^{Q} n_k \beta_{kj}^2} \sum_{q=1}^{Q} n_q \\
&\qquad \cdot \int_{|x-\alpha_{qj}|>\tau B_{nj}} (x-\alpha_{qj})^2 dF_{qj}(x) \\
&= \lim_{n\to\infty} \sum_{q=1}^{Q} \frac{n_q}{\sum_{k=1}^{Q} n_k \beta_{kj}^2} \\
&\qquad \cdot \int_{|x-\alpha_{qj}|>\tau B_{nj}} (x-\alpha_{qj})^2 dF_{qj}(x) \\
&\leq \lim_{n\to\infty} \sum_{q=1}^{Q} \frac{1}{\beta_{qj}^2} \int_{|x-\alpha_{qj}|>\tau B_{nj}} (x-\alpha_{qj})^2 dF_{qj}(x) = 0.
\end{aligned}$$

Since each variance β_{qj}^2 is positive and finite, and $B_{nj} = \sqrt{n_1\beta_{1j}^2 + \cdots, +n_Q\beta_{Qj}^2} \to \infty$ as at least one $n_q \to \infty$, each integral converges to zero as $n \to \infty$, and the Lindeberg condition is verified. Q.E.D.

Remark: The assumption that the $X^{(i)}$, $i = 1, 2, \ldots$, are generated from a finite number of normal distributions is stronger than necessary. This proof holds if $X^{(i)}$, $i = 1, 2, \ldots$, are generated from a finite number of arbitrary distributions.

Experimental results match this theoretical result, as illustrated by one experiment in Fig. 3.

C. Association Significance Test

The hypothesis test used in IGSCR to assess the significance of a cluster association to a class is based on the normal approximation to the binomial distribution (4). The null hypothesis is that the true probability of a pixel belonging to the majority class (for the cluster of interest) is less than p_0, a user supplied value. If $P(Z > \hat{z}) < \alpha$, where α is the user provided Type-I error, then the null hypothesis is rejected. The null hypothesis corresponds to the case when the cluster is impure, and rejecting the null hypothesis equates with labeling the cluster pure; if the null hypothesis is *not* rejected, the cluster is impure and the cluster is "rejected."

The hypothesis test for pure clusters in CIGSCR is different as the Bernoulli trials are fixed and testing the probability p of a success is no longer relevant. A pure soft cluster should have large weights for the majority class and comparatively small

weights for other classes. One possible hypothesis test compares the average weight for one particular cth class with the overall average weight for all classes in the jth cluster. Starting with the normal approximation for the sum of the cluster weights, the standard normal test statistic would be

$$\hat{z} = \frac{\sum_{i \in J_c} \big(w_{ij} - \mathrm{E}[W_{ij}]\big)}{\sqrt{\sum_{i \in J_c} \mathrm{Var}[W_{ij}]}},$$

where J_c is the index set of pixels prelabeled with the cth class. $\mathrm{E}[W_{ij}]$ and $\mathrm{Var}[W_{ij}]$ are unknown, but can be reasonably approximated using the sample mean

$$\overline{w}_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}$$

and sample standard deviation

$$S_{\overline{w}_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (w_{ij} - \overline{w}_j)^2}.$$

The Wald statistic is then

$$\hat{z} = \frac{\sqrt{n_c}(\overline{w}_{c,j} - \overline{w}_j)}{S_{\overline{w}_j}}, \tag{9}$$

where $n_c = |J_c|$ and

$$\overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}.$$

Since $\hat{z}$ is generated (approximately) by the standard normal distribution, a hypothesis test can be formed where the null hypothesis is that the average cluster weights corresponding to the cth class *are not* significantly different from the average cluster weights corresponding to all classes, and the alternate hypothesis is that the average cluster weights corresponding to the cth class *are* significantly different from the average cluster weights corresponding to all classes. Again, since class memberships are known a priori and all pixels have some positive membership with all clusters, testing for class memberships is not meaningful, but testing for significantly different cluster weights is meaningful. If $P(Z > \hat{z}) < \alpha$, the probability of observing the difference in the average cluster weights associated with c and the average cluster weights associated with all classes in the jth cluster is significant, and the null hypothesis is rejected. If the null hypothesis is *not* rejected, the cluster itself is rejected as impure, and further refinement is necessary.

One potential issue with the above test is that the sample mean and standard deviation calculations assume the sample is identically distributed, which is specifically *not* the assumption in this case. A better hypothesis test acknowledges that the data are not identically distributed, but are generated from a finite number of distributions. Since the number of distributions and the distributions are unknown, the number of classes and the individual class labels, which are assumed to correspond to inherent structure of the data, are used to approximate the true mean and variance of multiple clusters. Precisely, assume that all labeled pixel indices i with distribution index $\psi(i) = q$ correspond to the same class label $\phi(i) = c$. If $i \in \psi^{-1}(q)$, then $i \in \phi^{-1}(c)$, but $i \in \phi^{-1}(c)$ does not imply $i \in \psi^{-1}(q)$ (more than one distribution can map to one class), and $J_c = \phi^{-1}(c) = \{i \mid \phi(i) = c, 1 \leq i \leq n\}$. The above hypothesis test requires modification to use class information. In the previous test,

$$\sum_{i \in J_c} w_{ij} = \sum_{i=1}^{n} w_{ij} \delta_{\phi(i),c},$$

$$\hat{z} = \frac{\sum_{i=1}^{n} \big(w_{ij} \delta_{\phi(i),c} - \mathrm{E}[W_{ij} \delta_{\phi(i),c}]\big)}{\sqrt{\sum_{i=1}^{n} \mathrm{Var}[W_{ij} \delta_{\phi(i),c}]}},$$

$$\sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - \mathrm{E}[W_{ij}\delta_{\phi(i),c}])$$
$$= \sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - a_{ij}\delta_{\phi(i),c})$$
$$= \sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - \alpha_{qj}\delta_{\phi(i),c}),$$

recalling that $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ for $i \in I_q$. Assume when $\phi(i) = c$, and distribution index $q = \psi(i)$ corresponds to $c = \phi(i)$, then α_{qj} can be approximated by γ_{cj}, the mean of class $c = \phi(i)$. Ideally α_{qj} should be approximated directly, but there is no way to know $\psi^{-1}(q)$, so essentially $\psi^{-1}(q) \subset \phi^{-1}(c)$ is being approximated by $\phi^{-1}(c)$. Unfortunately, using the sample mean of the cth class and the jth cluster to approximate γ_{cj} and therefore α_{qj} breaks down because the sample mean of the cth class and the jth cluster is both the random variable on the left side and the approximation of the expected value on the right side of the minus sign. This is illustrated below. Approximating γ_{cj} (and α_{qj}) with the sample mean for the cth class,

$$\gamma_{cj} \approx \overline{w}_{c,j} = \frac{\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\sum_{k=1}^{n}\delta_{\phi(k),c}},$$

the numerator of the test statistic $\hat{z}$ becomes

$$\sum_{i=1}^{n}\left(w_{ij}\delta_{\phi(i),c} - \overline{w}_{c,j}\delta_{\phi(i),c}\right)$$
$$= \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \frac{\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\sum_{k=1}^{n}\delta_{\phi(k),c}}\sum_{i=1}^{n}\delta_{\phi(i),c}$$
$$= \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c} = 0.$$

Thus this test statistic does not work because the value being tested is the same as the estimated mean for the cth class when using the Kronecker delta instead of Bernoulli random variables. Recall that $Y_{c,j} = \sum_{i=1}^{n} V_{ic}W_{ij}$, where V_{ic}, $i = 1, \ldots, n$ are known prior to classification/clustering. Consider now the test statistic

$$\hat{z} = \frac{y_{c,j} - \mathrm{E}[Y_{c,j}]}{\sqrt{\mathrm{Var}[Y_{c,j}]}}.$$

Fixing j and c, and recalling that $n_q = |I_q|$, the number of indices i for which $X^{(i)}$ has the qth distribution,

$$\mathrm{E}[Y_{c,j}] = \mathrm{E}\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] = \sum_{i=1}^{n}\mathrm{E}[W_{ij}V_{ic}]$$
$$= \sum_{i=1}^{n}\mathrm{E}[W_{ij}]\mathrm{E}[V_{ic}] = \sum_{q=1}^{Q} n_q\alpha_{qj}p_c$$
$$= p_c\sum_{q=1}^{Q} n_q\alpha_{qj},$$

where p_c is the probability that $V_{ic} = 1$. Assuming all the pixels are independent and recalling that $\text{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$ where $i \in I_q$,

$$\begin{aligned}
\text{Var}[Y_{c,j}] &= \text{Var}\left[\sum_{i=1}^{n} W_{ij} V_{ic}\right] = \sum_{i=1}^{n} \text{Var}[W_{ij} V_{ic}] \\
&= \sum_{i=1}^{n} \big(\text{E}[W_{ij}^2 V_{ic}^2] - \text{E}[W_{ij} V_{ic}]^2\big) \\
&= \sum_{i=1}^{n} (p_c \text{E}[W_{ij}^2] - p_c^2 a_{ij}^2) \\
&= \sum_{i=1}^{n} (p_c (b_{ij}^2 + a_{ij}^2) - p_c^2 a_{ij}^2) \\
&= \sum_{q=1}^{Q} n_q \big(p_c (\beta_{qj}^2 + \alpha_{qj}^2) - p_c^2 \alpha_{qj}^2\big) \\
&= p_c \sum_{q=1}^{Q} n_q (\beta_{qj}^2 + (1 - p_c)\alpha_{qj}^2).
\end{aligned}$$

In the above formula, p_c would be approximated by its maximum likelihood estimate $n_c/n = |J_c|/n$. In order to estimate α_{qj}, assume that the qth distribution corresponds to the cth class, $\psi^{-1}(q) \subset \phi^{-1}(c)$, and

$$\alpha_{qj} \approx \overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}, \quad c = 1, \ldots, C,$$

where C is the number of classes. Then

$$\begin{aligned}
\text{E}[Y_{c,j}] &= p_c \sum_{q=1}^{Q} n_q \alpha_{qj} \approx p_c \sum_{d=1}^{C} n_d \cdot \frac{1}{n_d} \sum_{i \in J_d} w_{ij} \\
&= \frac{n_c}{n} \sum_{i=1}^{n} w_{ij} = n_c \overline{w}_j,
\end{aligned}$$

and

$$\begin{aligned}
\text{Var}[Y_{c,j}] &= p_c \sum_{q=1}^{Q} n_q (\beta_{qj}^2 + (1 - p_c)\alpha_{qj}^2) \\
&\approx p_c \sum_{d=1}^{C} n_d (S_{\overline{w}_{d,j}}^2 + (1 - p_c)\overline{w}_{d,j}^2),
\end{aligned}$$

where

$$S_{\overline{w}_{d,j}}^2 = \frac{1}{n_d - 1} \sum_{i \in J_d} (w_{ij} - \overline{w}_{d,j})^2.$$

Using these expressions for the mean and variance of $Y_{c,j}$, the Wald statistic is

$$\hat{z} = \frac{y_{c,j} - n_c \overline{w}_j}{\sqrt{p_c \sum_{d=1}^{C} n_d \big(S_{\overline{w}_{d,j}}^2 + (1 - p_c)\overline{w}_{d,j}^2\big)}}, \tag{10}$$

and the null hypothesis is rejected if $P(Z > \hat{z}) < \alpha$.

VI. Iteration

Together with the cluster association significance test, the iteration forms the semisupervised clustering framework in CIGSCR. The application of a hypothesis test determines which clusters should be used for classification, and an iteration works to produce a set of associated clusters with each class being represented by at least one associated cluster. This is accomplished by introducing new clusters that are likely to be associated, and when necessary, are associated with a class not already represented by a cluster.

In IGSCR, pure hard clusters are removed from the image that is clustered in subsequent iterations, focusing further refinement on clusters that failed to pass the purity test. K clusters are used for each iteration, presumably producing smaller clusters as less data is divided into the same number of clusters. The underlying assumption is that clusters that fail to pass the purity test could actually be composed of multiple clusters that would pass the purity test individually, and clustering the remaining data into K more clusters will reveal these smaller clusters. This method will not directly work on soft clusters as soft clusters cannot be removed simply by removing any sample associated with a pure cluster—all samples have a positive probability of belonging to any particular cluster.

In CIGSCR, unassociated clusters are targeted for refinement by using their information to create new clusters that will likely be associated. IGSCR is effectively locating smaller clusters that when combined to form a larger cluster would have been rejected. IGSCR accomplishes this by finding the same number of clusters (K) in the original dataset and then in successively smaller subsets of that original dataset. A similar approach that would locate smaller pure clusters in rejected clusters is "splitting" a cluster, employed by Ball and Hall [37] in ISODATA. Clusters are split by partitioning a cluster into two new clusters and recalculating new means. Soft clusters are represented by cluster means, and splitting a soft cluster would equate with replacing one cluster mean with two cluster means (calculated based on data associated with a cluster).

A cleaner algorithmic solution is to add one new cluster using information contained in the target cluster (the cluster that would be split), which effectively splits the cluster into two clusters. When using a clustering algorithm based on objective function (6), adding a new cluster guarantees a smaller function value (shown below) when $p = 2$. Using only the labeled samples belonging to the majority class (as determined in the cluster association significance test) to seed a new cluster would have the effect of pulling the new cluster toward those samples. Once another clustering iteration is completed, the targeted cluster would produce one cluster that is likely to be associated with the majority class and another cluster that retains relatively strong associations with all other classes. In CIGSCR, once the association significance test is performed, if at least one cluster is unassociated (and there are no unassociated classes), the cluster with the lowest value of $\hat{z}$ is used to generate a new cluster. The new cluster mean is determined using

$$U^{(K+1)} = \frac{\sum_{i \in \phi^{-1}(c_k)} w_{ik} X^{(i)}}{\sum_{i \in \phi^{-1}(c_k)} w_{ik}}, \tag{11}$$

where k is the cluster with the lowest value of $\hat{z}$, c_k is the majority class in cluster k, and recall that $\phi^{-1}(c)$ is the index set of labeled samples whose label is c. This formula also works when a class other than the majority class is used to seed a new cluster mean.

A shortcoming in IGSCR is that there is no guarantee that any clusters will be created and labeled with any particular class, and if a particular class is not represented by a cluster, the desired classification cannot be performed. In CIGSCR, this issue is addressed by adding a new cluster using information from a particular class if that class is not represented in the associated clusters. If a class c is not represented in the associated clusters, the cluster that is closest to being associated with c is used to generate a new cluster using (11) with $c_k = c$. The "closest" cluster is determined to be the cluster with the highest ratio of the average membership of class c to the average membership of the majority class.

When there are classes not represented by associated clusters and there are unassociated clusters, only one method can be used to determine the creation of a new cluster. If a cluster is unassociated, it is simply not used in classification. It is more important to have each class represented by the associated clusters than to refine an unassociated cluster, because the desired classification cannot be applied unless all classes are represented by associated clusters. Therefore adding a new cluster so that all classes will be represented takes precedence over adding a new cluster because an existing cluster is unassociated.

Finally, the theorem proving that adding one cluster mean will result in a smaller value of (6) is presented below.

Theorem: Given an integer $K > 0$, positive real numbers ρ_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, K+1$, defining a point $\rho \in \Re^{n \times K+1}$, and the objective function

$$J^{(K)}(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^2 \rho_{ij},$$

for K clusters where

$$w_{ij} = \frac{1/\rho_{ij}}{\sum_{k=1}^{K} 1/\rho_{ik}},$$

the objective function

$$J^{(K+1)}(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K+1} \hat{w}_{ij}^2 \rho_{ij},$$

for $K+1$ clusters where

$$\hat{w}_{ij} = \frac{1/\rho_{ij}}{\sum_{k=1}^{K+1} 1/\rho_{ik}},$$

satisfies

$$J^{(K+1)}(\rho) < J^{(K)}(\rho).$$

Proof: Note that the ρ_{ij} do not change with the addition of the $(K+1)$st cluster prototype, however $\hat{w}_{ij} < w_{ij}$ for $j < K+1$ because the denominator of $\hat{w}_{ij}$ has an additional term. Let $J_i^{(K)} = \sum_{j=1}^{K} w_{ij}^2 \rho_{ij}$ and $J_i^{(K+1)} = \sum_{j=1}^{K+1} \hat{w}_{ij}^2 \rho_{ij}$. It is sufficient to show that $J_i^{(K+1)} < J_i^{(K)}$ for each i to prove that $J^{(K+1)} < J^{(K)}$.

Let

$$S_1 = \sum_{k=1}^{K} 1/\rho_{ik} \qquad \text{and} \qquad S_2 = \sum_{k=1}^{K+1} 1/\rho_{ik}.$$

Then

$$w_{ij}^2 = \frac{(1/\rho_{ij})^2}{S_1^2} \qquad \text{and} \qquad \hat{w}_{ij}^2 = \frac{(1/\rho_{ij})^2}{S_2^2}.$$

$$\begin{aligned} J_i^{(K)} - J_i^{(K+1)} &= \sum_{j=1}^{K} \frac{(1/\rho_{ij})}{S_1^2} - \sum_{j=1}^{K+1} \frac{(1/\rho_{ij})}{S_2^2} \\ &= \frac{S_2^2 \sum_{j=1}^{K} (1/\rho_{ij}) - S_1^2 \sum_{j=1}^{K+1} (1/\rho_{ij})}{S_1^2 S_2^2}. \end{aligned}$$

Examining only the numerator in the previous term,

$$\begin{aligned} &(S_1 + (1/\rho_{i,K+1}))^2 \sum_{j=1}^{K} (1/\rho_{ij}) \\ &\qquad - S_1^2 \left(\sum_{j=1}^{K} (1/\rho_{ij}) + (1/\rho_{i,K+1}) \right) \\ &= (S_1 + (1/\rho_{i,K+1}))^2 S_1 - S_1^2 (S_1 + (1/\rho_{i,K+1})) \\ &= S_1^3 + 2S_1^2 (1/\rho_{i,K+1}) + S_1 (1/\rho_{i,K+1})^2 \\ &\qquad - S_1^3 - S_1^2 (1/\rho_{i,K+1}) \\ &= S_1^2 (1/\rho_{i,K+1}) + S_1 (1/\rho_{i,K+1})^2 \\ &> 0 \end{aligned}$$

yielding

$$J_i^{(K+1)} < J_i^{(K)}.$$

Q.E.D.

Assuming that the clustering algorithm locates a local minimum point of the objective function, the combination of the clustering algorithm and this cluster prototype addition are guaranteed to move toward a smaller objective function value. If

left unchecked, infinitely many clusters could be added, and the algorithm would continue to find smaller objective function values. The association significance test plays a crucial role in the termination of this iterative process. Once all clusters pass the association significance test and each class has at least one associated cluster, the iteration stops because the higher level objective has been met: clusters that significantly correspond to all classes have been located. The iteration also terminates when a maximum number of clusters is reached, and only those clusters that pass the association significance test are used for classification.

VII. Conclusions

This paper introduced a hypothesis test that can be used to evaluate the suitability of soft clusters for classification and suggested an iteration scheme that can be used to refine soft clusters. This hypothesis test was based on a normal approximation to a sum of random variables, and this approximation was proved reasonable under certain assumptions. This paper also provided a proof that the proposed soft cluster iterative refinement scheme will improve an objective function value when the soft k-means clustering algorithm is used. This association significance test and iteration will be necessary to convert IGSCR to use soft clustering to produce soft classifications. CIGSCR, the classification algorithm that incorporates the soft cluster evaluation and refinement presented here, is described in detail in Part 2. Part 2 also provides experimental results for IGSCR and CIGSCR.

References.

[1] Z. Sha, Y. Bai, Y. Xie, M. Yu, and L. Zhang, "Using a hybrid fuzzy classifier (HFC) to map typical grassland vegetation in Xilin River Basin, Inner Mongolia, China," *International Journal of Remote Sensing*, vol. 29, pp. 2317–2337, 2008

[2] A. Kumar, S.K. Ghosh, and V.K. Dadhwal, "Full fuzzy land cover mapping using remote sensing data based on fuzzy c-means and density estimation," *Canadian Journal of Remote Sensing*, vol. 33, pp. 81–87, 2007

[3] M. Pepe, L. Boschetti, P.A. Brivio, and A. Rampini, "Accuracy benefits of a fuzzy classifier in remote sensing data classification of snow," *Proc. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '07), VOLS 1–4*, pp. 492–497, 2007

[4] F. Okeke and A. Karnieli, "Methods for fuzzy classification and accuracy assessment of historical aerial photographs for vegetation change analyses. Part I: algorithm development," *International Journal of Remote Sensing*, vol. 27, pp. 153–176, 2006

[5] D.E. Sabol, J.B. Adams, and M.O Smith, "Quantitative subpixel spectral detection of targets in multispectral images," *Journal of Geophysical Research — Planets*, vol. 97 no. E2, pp. 2659–2672, 1992

[6] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. IT-14, pp. 55–63, 1968

[7] B.M. Shahshahani and D.A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes Phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, 1994

[8] J.P. Wayman, R.H. Wynne, J.A. Scrivani, and G.A. Reams, "Landsat TM-based forest area estimation using Iterative Guided Spectral Class Rejection," *Photogrammetric Engineering & Remote Sensing*, vol. 67, pp. 1155–1166, 2001

[9] R.F. Musy, R.H. Wynne, C.E. Blinn, J.A. Scrivani, and R.E. McRoberts, "Automated Forest Area Estimation via Iterative Guided Spectral Class Rejection," *Photogrammetric Engineering & Remote Sensing*, vol. 72, pp. 949–960, 2006

[10] R.D. Phillips, L.T. Watson, and R.H. Wynne, "Hybrid image classification and parameter selection using a shared memory parallel algorithm," *Computers & Geosciences*, vol. 33, no. 7, pp. 875–897, 2007

[11] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006

[12] M. Seeger, "Learning with labeled and unlabeled data," technical report, Inst. Adaptive and Neural Computation, University of Edinburgh, Edinburgh, UK, TR. 2001, 2001

[13] M. Belkin and P. Niyogi, "Semisupervised learning on Riemannian manifolds," *Machine Learning*, vol. 56, no. 1-3, pp. 209–239, 2004

[14] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means clustering with background knowledge," *Proc. of the 18th International Conference on Machine Learning (ICML '01)*, pp. 577-584, 2001

[15] I. Davidson and S.S. Ravi, "The complexity of non-hierarchical clustering with instance and cluster level constraints," *Data Mining and Knowledge Discovery*, vol. 14, no. 1, pp. 25–61, 2007

[16] S. Basu, A. Banerjee, and R.J. Mooney, "Active semi-supervision for pairwise constrained clustering," *Proc. of the SIAM International Conference on Data Mining*, 2004

[17] M. Bilenko, S. Basu, and R.J. Mooney, "Integrating constraints and metric learning in semisupervised clustering," *Proc. of the 21st International Conference on Machine Learning (ICML '04)*, 2004

[18] M. Halkidi, D. Gunopulos, M. Vazirgiannis, N. Kumar, and C. Domeniconi, "A clustering framework based on subjective and objective validity criteria," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 4, article 18, 25pp., 2008

[19] A. Bouchachia and W. Pedrycz, "Data clustering with partial supervision," *Data Mining and Knowledge Discovery*, vol. 12, no. 1, pp. 47–78, 2006

[20] N. Tishby, F.C. Pereira, and W. Bialek, "The informational bottleneck method," *Proc. of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999

[21] N. Slonim and N. Tishby, "Agglomerative information bottleneck," *Proc. of Neural Information Processing Systems Conference (NIPS '99)*, pp. 617–623, 1999

[22] S. Still, W. Bialek, and L. Bottou, "Geometric clustering using the information bottleneck method," *Proc. of Neural Information Processing Systems Conference (NIPS '03)*, 2003

[23] S. Kaski, J. Sinkkonen, and A. Klami, "Discriminative clustering," *Neurocomputing*, vol. 69, no. 1–3, pp. 18–41, 2005

[24] B. Jeon and D.A. Landgrebe, "Partially supervised classification using weighted unsupervised clustering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 1073–1079, 1999
[25] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVMs for semi- supervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363-3373, 2006
[26] G. Camps-Valls, T.V. Bandos Marsheva, and D. Zhou, "Semisupervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044-3054, 2007
[27] L. Gómez-Chova, L. Bruzzone, G. Gamps-Valls, and J. Calpe-Maravilla, "Semisupervised remote sensing image classification based on clustering and kernel means," *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS '08)*, 2008
[28] H. Jiang, J.R. Strittholt, P.A. Frost, and N.C. Slosser, "The classification of late seral forests in the Pacific Northwest, USA using Landsat ERM+ imagery," *Remote Sensing of Environment*, vol. 91, no. 3–4, pp. 320–331, 2004
[29] M. Kelly, D. Shaari, Q.H. Guo, and D.S. Liu, "A comparison of standard and hybrid classifier methods for mapping hardwood mortality in areas affected by "sudden oak death"," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 11, pp. 1229–1239, 2004
[30] R. Sivanpillai, C.T. Smith, R. Srinivasan, M.G. Messina, and X. Ben Wu, "Estimating regional forest cover in East Texas using enhanced thematic mapper (ETM plus) data," *Forest Ecology and Management*, vol. 218, no. 1–3, pp. 342–352, 2005
[31] R.H. Wynne, K.A. Joseph, J.O. Browder, and P.M. Summers, "Comparing farmer-based and satellite-derived deforestation estimates in the Amazon basin using a hybrid classifier," *International Journal of Remote Sensing*, vol. 28, no. 6, pp. 1299–1315, 2007
[32] B.V. Gnedenko, *Theory of Probability (sixth ed.)*, Gordan and Breach Science Publishers, The Netherlands, 1997, 497pp
[33] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007, 466 pp
[34] J.A. Richards and X. Jia, *Remote Sensing Digital Image Analysis (third ed.)*, Springer-Verlag, Berlin, 1999, 363 pp
[35] J. Bezdek, "Fuzzy mathematics in pattern classification," PhD Thesis, Cornell University, Ithaca, NY, 1974
[36] J.C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 1, pp. 1–8, 1980
[37] Ball, G.H. and Hall, D.J., 1965, "A novel method of data analysis and pattern classification," Stanford Research Institute, Menlo Park, CA